NAVAL POSTGRADUATE SCHOOL
Monterey, California

THESIS

VISUAL FEEDBACK FOR A STUDENT LEARNING LANGUAGE PRONUNCIATION

by

Kenneth H. Fritzsche

September 1997

Advisor: Nelson D. Ludlow

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.

1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 1997	**3. REPORT TYPE AND DATES COVERED** Master's Thesis

4. TITLE AND SUBTITLE VISUAL FEEDBACK FOR A STUDENT LEARNING LANGUAGE PRONUNCIATION	**5. FUNDING NUMBERS**
6. AUTHOR(S) Fritzsche, Kenneth H.	
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**

11. SUPPLEMENTARY NOTES

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release; distribution unlimited.	**12b. DISTRIBUTION CODE**

13. ABSTRACT ***(maximum 200 words)***

The learning of language pronunciation can be a frustrating and time-consuming process. Traditional methods require feedback from another person, usually an instructor or another student, or use of a self-assessment technique such as the listen-record-and- compare technique. These techniques have flaws. Human factors such as self confidence, shyness, fatigue, hearing ability, vocal tract agility, and confidence in the instructors fairness and competence all influence how rapidly a student acquires new pronunciation skills. A new technique to replace or augment existing techniques needs to be explored.

This thesis proposes the use of a computer to provide visual feedback to both complement auditory feedback to a student and lessen the detrimental impact that these human factors have on learning pronunciation. A computer shows no bias and provides an environment that affords privacy and the ability to practice whenever the student is willing. Additionally, the use of a computer to provide visual feedback helps a student to better understand exactly what portions and in what areas the pronunciation attempt is incorrect.

This thesis identified three required areas of pronunciation feedback - phoneme, stress and intonation, and integrated them into a single interface. An object-oriented LISP implementation is presented to display the visual feedback and a design for digital speech processing is proposed to analyze the pronunciation and supply the interface with data.

14. SUBJECT TERMS Visual Feedback, Pronunciation	**15. NUMBER OF PAGES** 80
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UL

VISUAL FEEDBACK FOR A STUDENT LEARNING LANGUAGE PRONUNCIATION

Kenneth H. Fritzsche
Captain, United States Army
B.S., United States Military Academy, 1988

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

from the

NAVAL POSTGRADUATE SCHOOL
September 1997

Author Kenneth H Fritzsche
Kenneth H. Fritzsche

Approved by: Nelson D. Ludlow
Nelson D. Ludlow, Thesis Advisor

Wayne J. Redenbarger, Second Reader

Timothy J Shimeall for
Ted Lewis, Chairman
Department of Computer Science

ABSTRACT

The learning of language pronunciation can be a frustrating and time-consuming process. Traditional methods require feedback from another person, usually an instructor or another student, or use of a self-assessment technique such as the listen-record-and-compare technique. These techniques have flaws. Human factors such as self confidence, shyness, fatigue, hearing ability, vocal tract agility, and confidence in the instructors fairness and competence all influence how rapidly a student acquires new pronunciation skills. A new technique to replace or augment existing techniques needs to be explored.

This thesis proposes the use of a computer to provide visual feedback to both complement auditory feedback to a student and lessen the detrimental impact that these human factors have on learning pronunciation. A computer shows no bias and provides an environment that affords privacy and the ability to practice whenever the student is willing. Additionally, the use of a computer to provide visual feedback helps a student to better understand exactly what portions and in what areas the pronunciation attempt is incorrect.

This thesis identified three required areas of pronunciation feedback - phoneme, stress and intonation, and integrated them into a single interface. An object-oriented LISP implementation is presented to display the visual feedback and a design for digital speech processing is proposed to analyze the pronunciation and supply the interface with data.

TABLE OF CONTENTS

I. INTRODUCTION

A. GOALS

The primary goal of this thesis is to see if it is possible to design an interface that provides useful meaningful data for a person attempting to learn pronunciation of a language. The main idea is to compare a test speech digital recording with a pre-analyzed "desired" segment of speech to determine errors in pronunciation and provide some form of immediate visual feedback for a student attempting to learn proper pronunciation. In order to accomplish this, some means of digital signal processing and comparison must be conducted to supply the interface with data.

B. BACKGROUND AND MOTIVATION

Language learning necessitates a student to acquire new knowledge in many areas at once. Some of this knowledge is abstract in nature and some physical. The abstract portion includes learning a new vocabulary, learning to write and recognize a new alphabet, and learning new rules for forming phrases and sentences. The physical side refers to a student learning how to pronounce new sounds not found in their native tongue and/or combining familiar sounds in unfamiliar combinations. It is aiding in this physical side of language acquisition that the research of this thesis attempts to explore.

Traditionally, learning language pronunciation has two major forms. The first requires that a student interacts with another person, preferably the teacher, and receives immediate feedback on how well they fared at their pronunciation attempt and what exactly were the identified errors. A second approach is the record-listen-and-compare technique. This involves a student listening to a prerecorded speech segment and recording his or her own attempt at pronunciation. Once recording of all attempts is complete, the student then goes back and compares their attempt to the prerecorded speech segments. It now falls on the shoulders of the student to evaluate their own performance. Both of these methods have their merits but both also contain potential flaws introduced by human error. Both methods are susceptible to student frustration, fatigue or inability to hear the difference, even when explicitly pointed out by the teacher. In the first technique, factors such as instructor fatigue, frustration, hearing problems or incorrect knowledge can lead to erroneous feedback. In the second technique, feedback is totally dependent of the student whose lack of knowledge or lack of ability to discern subtle differences may introduce erroneous self-feedback. Additionally, in the second method the feedback is not immediate as it requires waiting until the recording process is complete and some fashion of resetting the taping device to play both speech segments. Another method, either in lieu of or in addition to these two methods, is needed to augment these two methods and to make up for where these two methods lack.

A method that uses a computer to provide immediate visual and acoustical feedback may help in this area. By immediately identifying errors and providing feedback, this method mirrors the student/instructor method described above. However, computers do not fatigue, get frustrated or have day-to-day hearing fluctuations such as when an instructor has a cold or has been subjected to a loud noise. By providing visual feedback, the student not only can hear the difference but also see the difference. This may help in cases where students may have difficulty hearing subtle differences. Finally, since a student knows that a computer is unbiased and consistent, perhaps a different level of confidence can be achieved on the part of the student trusting that the feedback is genuine and not due to instructor partiality or fatigue.

C. RESEARCH QUESTIONS

This thesis will examine the following research areas:

- How is speech produced and how can it be classified and represented?
- What causes pronunciation errors and in what areas might a student attempting to learn proper pronunciation make errors?
- What areas must an interface address to provide useful visual feedback to a student attempting to learn a foreign language?
- Can this interface adequately integrate and represent pronunciation errors?
- What is a possible approach using speech processing (digital signal processing) techniques to identify errors and supply the interface useful data?

D. SCOPE

The potential areas encompassed by this thesis are huge and range from design of graphical interfaces, to speech and linguistic study, to learning pronunciation rules and nuances of different foreign languages, to advanced digital signal processing. To keep the area of study manageable, limitations had to be set. While delving into the deep abyss of speech recognition and signal processing it was desired to keep with a familiar language, English, in developing the speech recognition engine. I decided early that primary focus was given to exploring the interface content and design, with minimal efforts going to exploring the speech processing engine, recognizing that current state of the art speech recognition technology can not yet meet the expectations required for this thesis.

E. ORGANIZATION

Chapter II of this thesis covers general information on how speech is produced and general speech recognition requirements and techniques. Chapter III discusses the methodology used to select the components of and design of the interface. Chapter IV proposes a design for the pronunciation feedback engine to supply the data to the interface and a LISP implementation of portions of the interface. Chapter V summarizes the thesis and presents the final results and conclusions on the components of the interface.

II. BACKGROUND

A. INTRODUCTION

In order to have a complete understanding of speech recognition one must first understand how humans produce the sounds that combine to form speech. By knowing how, where and when speech is produced it is possible to teach and correct proper pronunciation. Understanding these characteristics allows for the analysis of digital speech signals and speech recognition. This chapter will be dedicated to this pursuit.

B. SPEECH

1. Speech Production

Sound is the transmission of vibrations through a medium such as air. By causing molecules of air to vibrate, these vibrations move out and away from the source in waves, which is perceived by human ears as sound. Speech is merely a specialized form of sound produced by humans. Humans have evolved and learned to control different portions of the speech tract to form and manipulate sounds into a complex form of communication we call speech. The way the different sounds of speech is produced, changed and combined in the different areas of the vocal tract (Figure 2-1) is what makes speech unique amongst the different languages.

Speech starts when air is forced from the lungs through the larynx. The larynx's purpose is to act as a valve between the lungs and the mouth, which is especially important when eating. It consists of two small folds of ligament, sometimes called the vocal cords, extending from the front to the back on either side and are held together at the front. The size of the opening and the amount of air that passes through can be controlled by varying the size of the separation of the folds at the back. When in the closed position, air pressure builds up behind the folds until they are blown apart. This resulting separation causes a decrease in the pressure and allows the folds to close and once again build up pressure. This rhythmic opening and closing of the vocal folds causes the folds to vibrate and produce sound known as voiced speech. If the vocal folds are open and slightly apart, turbulence results in the airflow but no vibrations occur. This is know as aspiration and is exemplified by the sound produced for a whisper or the letter 'h' in the word "hasp". Finally, the vocal folds can be wide open, resulting in voiceless sound such as all of the consonants in the word **shafts**.

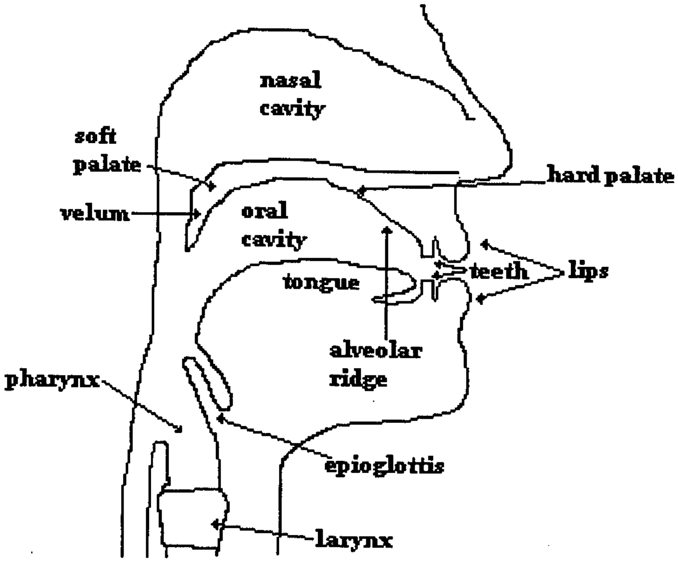


Figure 2-1 Vocal Tract

The frequency of the vocal fold vibration is determined by their mass, tension, length and air pressure forced through them. For any given speaker, the last three can be controlled. Typically ranges for vocal fold vibrations for humans is from 60-400 Hertz, with the average for an adult male being 100 Hz and the average for an adult female 180 Hertz [Ref. 1].

Once the air passes through the larynx, it then passes through the rest of the vocal tract whose shape and size at various locations can be modified to produce different sounds. The pharynx can be lengthened and narrowed, changing the sound characteristics. The velum can be raised or lowered, allowing air to either pass through both the nasal and oral cavities or through the oral cavity only, producing nasal and non-nasal sound. The hard palate, aveolar ridge, tongue, teeth and lips can be positioned to briefly stop or further disrupt the airflow, all affecting the type of sound produced. The location and manner in which the speech sound is produced allows the sound to be classified.

2. Classes of Speech Sounds

The human vocal tract is capable of forming two basic classes of sounds, consonants and vowels. Consonants are formed with a relatively closed vocal tract which at times may be completely closed. This closure or restriction of the vocal tract disrupts the air flow and produces and audible change in the sound. Vowels on the other hand, are produced with a relatively open vocal tract, with little or no disruption of the airflow.

The table below shows how consonants are classified based on the location in the vocal tract where they are produced.

Classification	*Explanation*	*Example*	*Location in Figure 2-2*
Labial	Main closure between the lips	**mom**	A
Labiodental	Lower lip approaches or closes against the upper teeth	**five**	A/B
Dental	Main closure between tongue and teeth	**thin**	B
Alveolar	Tongue approaches or closes against the alveolar ridge	**none**	C
Palatal	Tongue approaches or closes against the hard palate	**sash**	D
Velar	Back of the tongue closes against the soft palate or velum	**kick**	E
Glottal	Aspirated speech	**hasp**	F

Table 2-1 Consonant Classification Based on Location

An examination of Figure 2-2 shows the articulation location for consonants. This location is a component that makes each consonant unique and serves as a characteristic used to classify consonants.

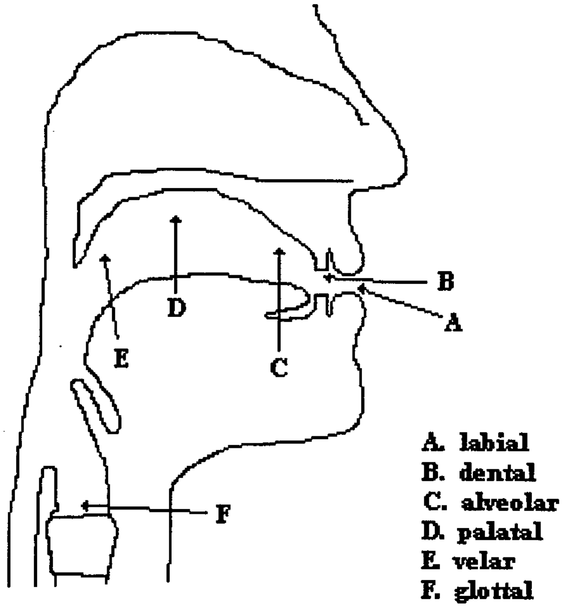


Figure 2-2 Articulation Location

In addition to location, the way in which sound is produced also serves to classify consonants. When there is a complete closure in the vocal tract, the air pressure builds up behind the closure and when released, there is a burst of energy like in the words boy and pig. These consonants are known as plosives. Nasals also have a complete closure in the vocal tract but unlike plosives, there is no build up of pressure as the velum is opened and the air is allowed to flow into the nasal cavity. An example of a nasal consonant can be seen in the word **mom**. Fricative consonants are formed when the vocal tract is constricted to the point where turbulence forms in front of the narrowed area. These can be seen in the word five. Affricatives are consonants that, like plosives and nasals, have a complete closure in the vocal tract, but differ because they have a much more gradual release resulting in a unique sound similar to fricatives. Examples of these are seen in the consonants in the words church and judge. The final classification of consonants is approximants. In these consonants, the vocal tract is constricted but does not result in fricative-like turbulence. The consonant is usually attached to a vowel and is produced initially in a vowel-like position. It then changes rapidly to the position for the adjoining vowel. Examples of approximants are web, you, rat, and let. These approximants are also referred to as semi-vowels, glides or liquids.

Like consonants, vowels are also classified according to the location of their articulation. The first consideration is the tongue height at the point of greatest point of closure. The words high, mid and low are used to describe this classification. Examples of high vowels are meat, bit, mid vowels are bet and bait, low vowels cat and bob. Next considered is the horizontal place of greatest closure of the mouth. The terms front, central and back are used to describe this. Examples of front vowels are meat, bit, and bait, central vowels but and bird, and back vowels cot, bought, caught, cook, and boot. Another way vowels are classified is by the roundness of the lips during formation. Rounded lips in English are present primarily in back vowels and in the vowels in the words pool, pull, and pole. The final way vowels are classifies is as diphthongs. A diphthong is a speech sound that begins with the vocal tract in the position and articulation of one vowel and then moving to another. Examples of this are the words boy and high.

3. The Phoneme

The phoneme is the smallest unit of speech sound that is discernible in a spoken language. American English consists of 40 phonemes, 16 vowels and 26 consonants. In order to study and analyze pronunciation in languages, phoneticians developed the International Phonetic Alphabet (IPA) in 1888 to describe all of the different phonemes in all the world's languages [Ref. 4]. The IPA consists of different symbols that are used to describe all of the phonemes and their possible variations. Consider the English transcription of the phrase "The International Phonetic Association":

ði ɪntə'næʃənəl fə'nɛtɪk əsoʊsi'eɪʃn

As can be seen, these IPA symbols are difficult to understand and awkward to use so the Advanced Research Projects Agency (ARPA) developed a single (common) symbol ARPAbet and later, an uppercase ARPAbet, to describe American English phonemes. These are described along with heir manner and location of articulation in Table 2-2. Throughout the remainder of this thesis the ARPAbet uppercase version will be used for any phoneme transcription.

It is important to now point out and recognize a difference that exists between the words phoneme and phone when describing speech sounds. A phoneme is an abstract unit, that is it is a theoretical concept used by phoneticians to discuss and differentiate between speech sounds. A phone is a physical manifestation of a phoneme. Its is the sound actually produced by a speaker. The true meaning of phone and phoneme have become blurred and often indistinguishable. Throughout this thesis the word phoneme will be applied in a more general sense and contextual clues will provide the true implied meaning of the word.

	Example	IPA Symbol	ARPAbet Single-Symbol	ARPAbet Uppercase	Manner	Place	Voiced
CONSONANTS	yes	j	y	Y	glide	front unrounded	yes
	with	w	w	W	glide	back rounded	yes
	like	l	l	L	liquid	alveolar	yes
	row	r	r	R	liquid	retroflex	yes
	mom	m	M	EM	nasal	labial	yes
	not	n	N	EN	nasal	alveolar	yes
	ring	ŋ	G	NX	nasal	velar	yes
	fair	f	f	F	fricative	labiodental	no
	valve	v	v	V	fricative	labiodental	no
	thing	θ	T	TH	fricative	dental	no
	them	ð	D	DH	fricative	dental	yes
	sass	s	s	S	fricative	alveolar	no
	zoos	z	z	Z	fricative	alveolar	yes
	show	ʃ	S	SH	fricative	palatal	no
	measure	ʒ	Z	ZH	fricative	palatal	yes
	hip	h	h	HH	fricative	glottal	no
	pot	p	p	P	stop	labial	no
	big	b	b	B	stop	labial	yes
	tip	t	t	T	stop	alveolar	no
	dip	d	d	D	stop	alveolar	yes
	kid	k	k	K	stop	velar	no
	gag	ɡ	g	G	stop	velar	yes
	church	ʧ	C	CH	affricate	palatal	no
	just	ʤ	J	JH	affricate	palatal	yes
VOWELS	feet	i	I	IY	vowel	high front	yes
	bid	I	I	IH	vowel	high front	yes
	wait	e	e	EY	vowel	high front	yes
	wet	ε	E	EH	vowel	high front	yes
	bat	æ	@	AE	vowel	low front	yes
	bog	ɑ	a	AA	vowel	low back	yes
	caught	ɔ	c	AO	vowel	mid back rounded	yes
	boat	o	o	OW	vowel	mid back rounded	yes
	hook	U	U	UH	vowel	high back rounded	yes
	hoot	u	u	UW	vowel	high back rounded	yes
	bud	ʌ	A	AH	vowel	mid back	yes
	burp	ɝ	R	ER	vowel	mid	yes
	ago	ə	x	AX	vowel	mid	yes
	kite	ɑI	Y	AY	dipthong	low back →high front	yes
	boy	ɔI	O	OY	dipthong	mid back → high front	yes
	about	ɑU	W	AW	dipthong	low back → high back	yes

Table 2-2 American English Phoneme Symbols and Classification

C. SPEECH DATA

Speech is the sound wave produced by the vocal tract of a person. It can be sensed in the ear but once it is produced and transmitted, it is gone. This section deals with how to store and represent speech for playback and analysis.

1. Analog Signals

Sound waves and thus speech is an analog signal. This means that for every point in time any time interval (say interval A-B in Figure 2-3), there is a corresponding value that represents the amplitude of that signal. Even though there is a maximum and minimum amplitude of the interval, there are an infinite number of values for the amplitude of that signal between these minimum and maximum values. Additionally, that interval can be broken into an infinite number of time values.

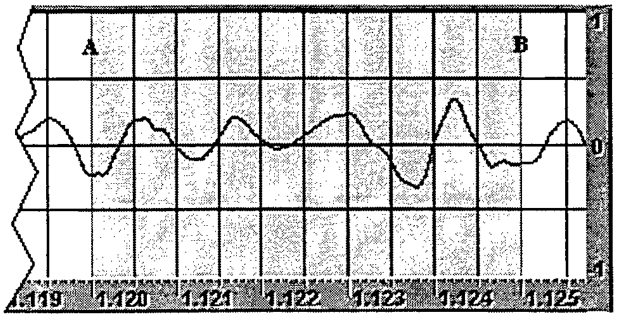


Figure 2-3 Analog Speech Signal Segment

It would be impossible to store an infinite number of amplitudes and an infinite number of time intervals in data form for use on a computer, so a finite digital representation of an analog signal is used.

2. Digital Signals

A digital signal represents an analog signal by storing amplitudes for specific time values or samples. An analog signal is converted to a digital signal in the following manner. First the time interval is broken into equally spaced time intervals. Then a set of values to represent possible signal amplitudes is determined. The process of breaking the signal into time intervals is known as sampling. For each sample, the closest in the set of possible amplitude values is determined and is mapped to that sample. This process is known as quantizing. The horizontal lines in Figure 2-4 depict these quantization levels, that is the set of possible values for amplitude.

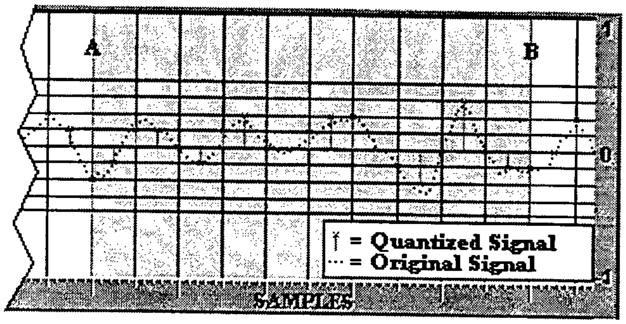


Figure 2-4 Digital Signal

The number of quantization levels determines the number of bits required to store the data for each sample and the how accurately the digital signal represents the analog signal. Typically, the minimum number of bits required for intelligible recording and playback of speech signals is 8 bits. For more advanced digital speech processing and speech recognition, 16 and sometimes even 24 bits per sample is used. The total amount of bits required to represent an analog signal as a digital signal is a function of the number of samples and the number of quantization levels. A large number of samples per second and a large number of quantization levels allow a very accurate digital representation of the analog signal. However, this tradeoff of having this large, accurate digital signal is the overhead costs of storage, and signal processing computations. Thus, it is desired to only have the minimum number of samples required to adequately represent the signal.

The Nyquist sampling theorem helps determine the minimum number of samples required. It states that an analog signal may be perfectly reconstructed if it is sampled at a rate greater than or equal to twice the frequency of the highest frequency component of the signal. For instance in music, which has a wide range of frequencies, the sampling rate standard is 44.1 kHz (Compact Disc) or 48.0 kHz (Digital Audio Tape). For speech, most vowels have relatively low frequency components, all below 4 kHz. This would make a sampling rate of 8 kHz (8000 samples in one second) more than adequate. However, certain consonants, fricatives, and female speech have energy in higher frequencies (in the range 6 - 8 kHz) and a sampling rate of 16 kHz is required to represent these speech signals for detailed analysis [Ref. 3].

3. The Speech Signal

A speech signal is a very complex, non-static signal. This means that the signal is actually made up of many different signals with different frequencies that change over time. Figures 2-5 and 2-6 illustrate the difference between a static single frequency signal and a speech signal. Figure 2-5 shows a static signal consisting of a single frequency of 100 Hz. This means that the wave will complete one cycle 100 times in a second. A cycle is characterized by the signal crossing the zero axis two times.

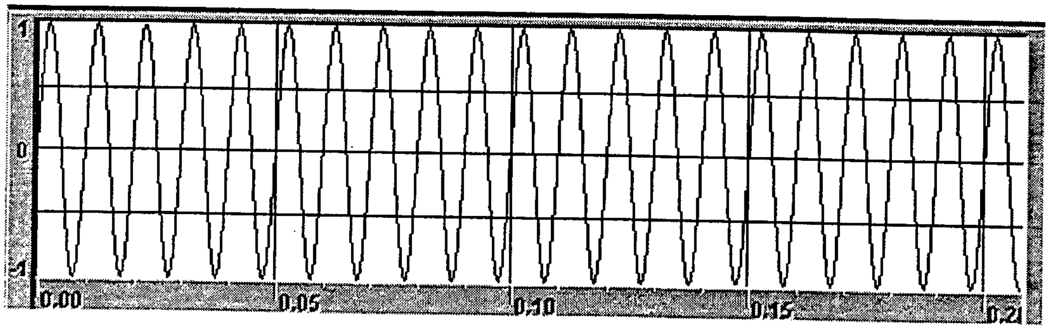


Figure 2-5 Static Signal, Frequency = 100 Hz

Figure 2-6 shows a non-static speech signal. Note how the shape and the size of the wave change over time. This is caused by different frequencies being present or not present at different times throughout the entire signal. These different frequencies are the result of what is produced by the different parts of the vocal tract, all combining to form a single speech signal that varies with time.

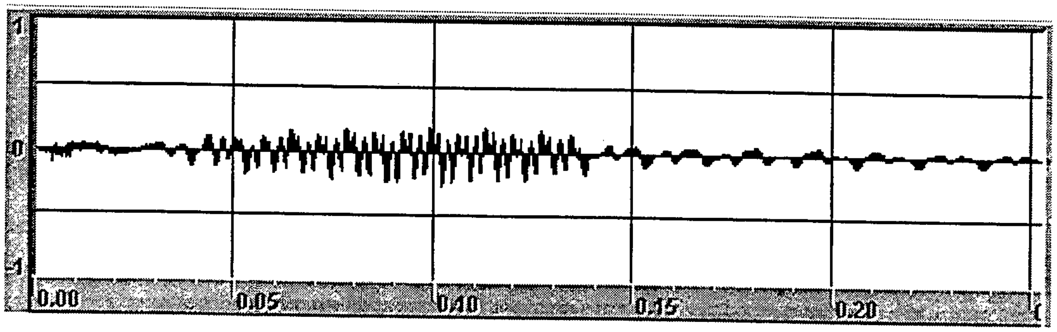


Figure 2-6 Speech Signal -The word "Good"

A quick examination of Figure 2-7 points out one of the difficulties in analyzing a speech signal, the fact that it is difficult to determine where exactly a word or a phoneme ends or begins. It appears that signal consists of there are five distinct wave segments that are characterized by periods of relatively higher amplitude. One would think that they correspond to either words or phonemes. This is not the case. In fact there are fifteen phonemes and six words in this signal. The labels in Figure 2-8 show the precise location of each phoneme. These segments of high amplitudes are actually the vowels. Since vowels are voiced, their amplitudes are larger. Examining signals by considering the amplitude over time is one form of time-domain analysis.

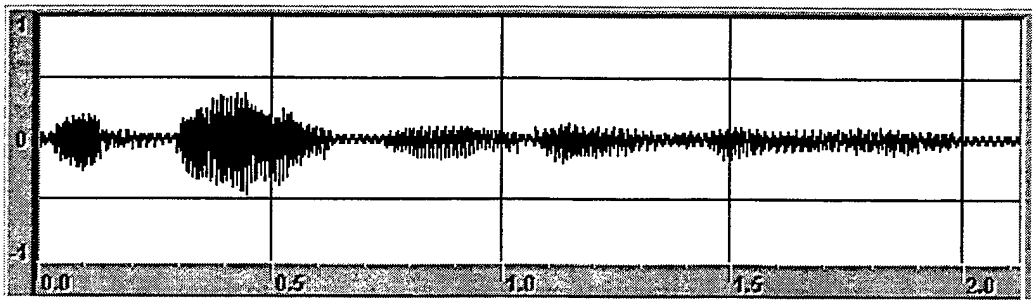


Figure 2-7 Speech Signal

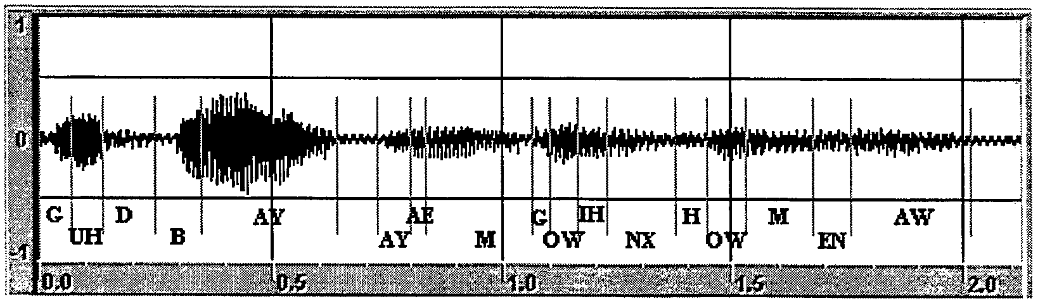


Figure 2-8 Speech Signal "Good bye, I am going home now".

Another time domain analysis technique is the zero crossing rate of a signal. The As mentioned previously, frequency is defined in cycles and a cycle is a signal crossing of the zero axis two times. The higher the frequency, the higher the zero crossing rate will be. Thus by considering the zero crossing rate it can be determined if the speech signal consists primarily of low or high frequency signals. If the speech segment has a low zero crossing rate then it is a cue that it is voiced segment. If it is high, then it is unvoiced. A common threshold between voiced and unvoiced speech is 2500 crossing/second.

A final use of time-domain analysis is to extract pitch. As previously stated, speech is formed of many different frequencies but there is usually a single fundamental frequency which can be measured. The time distance of this fundamental frequency's cycle is known as the pitch period.

Another area of speech signal analysis is known as frequency-domain analysis. This uses a common signal processing tool known as the Fast Fourier Transform. An FFT considers a speech signal over a specific interval and provides information on all of the frequency components present in a signal for that interval. It employs a technique referred to as windowing where the values of the amplitudes across the interval can be weighted in different fashions. Two common windows are Rectangular and the Hamming. In a Rectangular Window, all of the samples in the interval are weighted equally. In a Hamming Window, samples in the middle of the interval are weighted more heavily than those near the interval boundaries. Its shape is similar to a bell curve used in statistical normal distributions. The information provided by an FFT can be plotted over time in a graph know as a spectrogram. Regardless of the window shape used, the FFT and the spectrogram is a useful tool in determining the frequencies of what is know as formants.

4. Formant Frequencies

Although speech signals consist of numerous signals of different frequency, they commonly contain up to five predominant or "major" frequencies known as formant frequencies. The frequencies at which theses formants are located can be used as key characteristics for speech recognition. It has been determined that for speech recognition, only the first three formant locations are important. The other two contain no useful information for application in speech recognition. However, they are of importance in producing natural sounding computer generated speech. Figure 2-9 shows a spectrogram of the word bye.

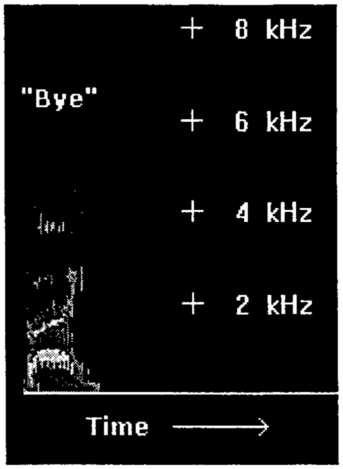


Figure 2-9 Spectrogram of the word "Bye".

The formant frequencies are portrayed by the most intense frequencies that are lighter in color. The first formant or F0 is located at between 800 and 1000 Hz. The second, F1, starts at about 1400 and rises to about 1700 Hz throughout the duration of the word. F3, F4 and F5 formant frequencies are located at around 3000, 3600 and 5000 Hz respectively. All speech phonemes do not necessarily contain five formants and sometimes two closely located formants can blend into one. Formant location can change abruptly, especially in transition to and from nasal phonemes. Thus, formant location identification for a given speech segment is a hard task in itself, and the task of tracking formants through time is extremely difficult.

D. EXISTING PRODUCTS

There are many multimedia language learning software packages on the market, all of which claim to assist in vocabulary building and in teaching pronunciation. These all have educational benefit and are useful to a point. Most display words or simple phrases and play recorded segments to teach pronunciation. Some even use the listen, record and compare technique. However, all fall short of the similar claims that many vendors use, describing their products as "like having your own personal language tutor" [Ref. 7] [Ref. 8] [Ref. 9].

These claims are false because none of them actually analyze and point out the pronunciation errors. Products for the PC that "use voice recognition to provide feedback" such as sold by The Learning Company and Syracuse Language Systems only provide feedback in the form of a needle gauge as shown in Figure 2-10 that indicates that a pronunciation attempt of a single word is somewhere on a scale between "tourist" and "native". These give the student no useful feedback on the specific area of mispronunciation, only some meaningless score that tells them that their attempt was between extremes.

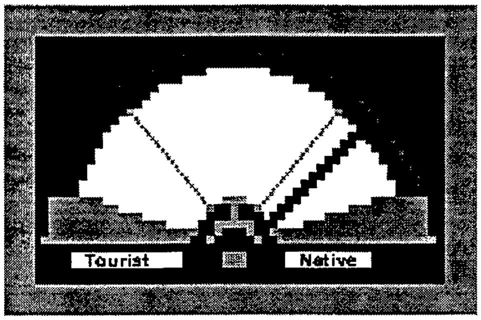


Figure 2-10 Gauge Style Feedback

Another product that uses speech recognition techniques to provide feedback is the Voice Interactive Language Training System (VILTS) by SRI International. VILTS is a workstation based system that requires a student to read a passage and fill in the blanks from a list of candidate words. The system then provides a pronunciation score "that correlate well with those of expert human listeners" [Ref. 6]. While a somewhat unique approach, this score is useless in providing students an indication on what specifically they are mispronouncing. A better system that provides meaningful information needs to be developed.

E. SUMMARY

The study of speech pronunciation requires an understanding of how speech is produced with air being forced from the lungs through the larynx and how the sound is modified by each of the major areas of the vocal tract until the final speech sound exits via the nose and mouth. It also requires an awareness that the basic unit of pronunciation is the phoneme and that phonemes combine to form words. It is at the phoneme level that students learn to pronounce new sounds when learning a new language. The process of automatic phoneme identification is a challenge due to coarticulation effects that cause a phoneme to vary based on adjacent phonemes.

Once speech is produced, it can then be recorded as a digital signal which requires a sampling rate and quantization levels which determine the amount of data that is needed to be stored to accurately represent the original speech signal. Once captured as a digital signal, the speech can be decomposed and analyzed to determine unique characteristics of the signal. Characteristics such as amplitude, frequency, zero crossing rate, and pitch all provide information to uniquely identify and compare speech segments. An important characteristic is the location of the formant frequencies or the predominant frequencies of a signal. Speech signals are made up of many different frequencies, all of which result from the different areas of the vocal tract and present in differing quantities. There are three to five formant frequencies present in a speech signal. A Fast Fourier Transform is a tool used to help identify the frequencies present in a given speech segment. A weighting factor or window is applied to this segment to focus the result of the FFT on specific areas of the segment. A spectrogram is a visual representation of how the FFT change with respect to time.

Existing language learning software that claim to provide feedback are of little use or meaningless. More specific and informative forms of feedback need to be used to give the student a clue on what they are doing wrong and what to change for better pronunciation. The next chapter will present a proposed process by which useful feedback can be collected and presented to a student.

III. INTERFACE DESIGN

A. SELECTING COMPONENTS OF A USEFUL INTERFACE

Determining and designing the layout of an interface to help a student identify pronunciation errors must address a few issues. First is what kind of errors can be made. Second, what information is required to help a student identify their errors. Lastly, in what manner can this information be presented.

1. Pronunciation Errors

Pronunciation errors can essentially be grouped into three general categories. First is improper pronunciation of phonemes. Second is the misapplication of stress to syllables. Last is the improper change in intonation in a syllable or word.

a. Improper Phoneme

Phoneme errors fall into three areas - mispronunciation of a phoneme, inclusion of a phoneme that does not belong or the omission of a phoneme that does belong.

(1) Phoneme Mispronunciation. In attempting to pronounce a word, a student may incorrectly reproduce one or more phonemes. This error may be caused by a student's inability to hear the difference between what they are saying and the correct pronunciation. Another source of the pronunciation error is the student's unfamiliarity in controlling the vocal tract to produce this sound either because it is not a sound present in the student's native language or because it is being combined with another phoneme resulting in an unfamiliar transition. Regardless of the cause, the result is still the same - an incorrect phoneme and thus an incorrect pronunciation.

(2) ***Phoneme Inclusion.*** Another possible form of an improper pronunciation attempt would be the inclusion of a phoneme or phonemes that are not present in the original word or phrase. This type of error could result from a student attempting to pronounce every letter that they see (such as the silent e in the English pronunciation of the word like) or attempting to pronounce the word or phrase using pronunciation rules of their native tongue.

(3) ***Phoneme Omission.*** The last possible form of an improper pronunciation attempt would be the omission of a phoneme or phonemes that is supposed to be present but is not pronounced by the student. This type of error could result from a student not being aware or using the wrong pronunciation rules. An example of this would be the popular sport shoe manufacturer Nike being pronounced as a one syllable word as opposed to the two syllables that is correct.

b. Stress on Wrong Syllable

In polysyllabic words, some syllables receive more stress than others do. This means that relative to the other syllables, certain syllables are pronounced louder than others and/or the duration of the syllable is longer. It is also common for words to have syllables with different level of stress. The word satisfactory (sat-is-fac-to-ry) has syllables of three different stress levels. The most stress is on the third syllable, the second most stress is on the first syllable and the remaining syllables receive equal levels of stress. Incorrect placement of stress on the wrong syllable can result in a spoken word or phase to be misunderstood or unintelligible.

c. *Wrong Inflection/Intonation*

Intonation changes, that is changes in the pitch or tone of a pronunciation, can provide clues as to what type of phrase is being spoken, whether it is a declarative (statement), an interrogative (question) or an imperative (command). Improper inflection can mislead a listener as to the nature of the phrase. Consider the declarative "The book is on the table." By raising the inflection of the last word, the meaning is changed from a declarative statement announcing the location of the book to an interrogative question as to whether the book is on the table. In tonal languages, the actual meaning of the word is determined exclusively by its inflection. For instance in Thai there are five different tonal pronunciations. First there is the low tone which is the voice pitched slightly lower than normal. Next there is the high tone which is the voice pitched slightly higher than normal. Next there is the rising tone similar to the question intonation in English. Next there is the falling tone which is pronounced as if you were emphasizing something or calling someone from afar. Finally there is the mid tone which is the normal speaking tone of an individual. Figure 3-1 illustrates the different tones in the Thai language.

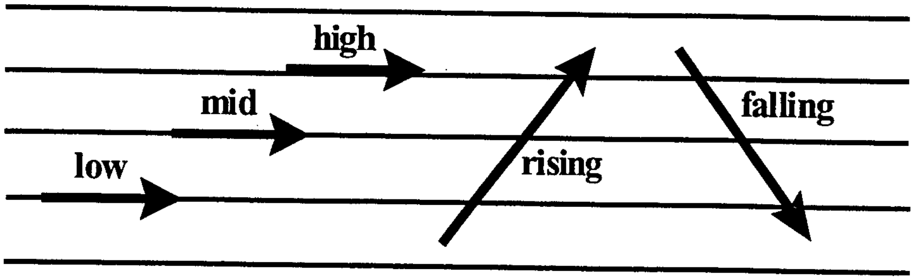


Figure 3-1 Different Types of Tone Pronunciation in Thai

The use of these different tones determines the meaning of the word. For instance the word mai has five meanings depending on the tone. Note the type of tone, meaning and tone symbol over the letter a for all the meanings of the word mai.

	Tone	Meaning
mái	high	new
mài	low	wood
mai	mid	not
mâi	falling	burns
mǎi	rising	not?

The expression mái mài mâi mai mǎi means "New wood doesn't burn, does it?" [Ref. 10]. Each word contains precisely the same phonemes. The only difference between words is the inflection used in their pronunciation. Thus, the wrong inflection on a word or phrase can have a dramatic affect on the meaning and must be addressed by an interface aimed at teaching proper pronunciation.

2. Level of Phrase Segmentation Required for Error Identification

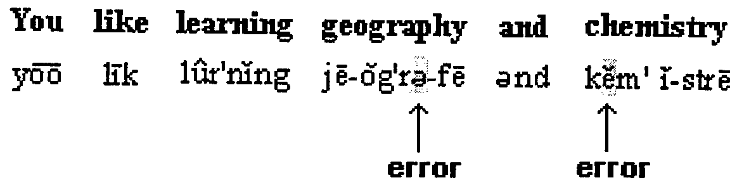
Since the basic unit of pronunciation is the phoneme, feedback must be down to the phoneme level. In order for a student to be able to understand precisely what portion of their attempt at correct pronunciation is wrong, the interface must show the specific phoneme or phonemes that are wrong. For stress, the phoneme may not be the ideal unit to express errors. Since stress is applied at the syllable level, some sort of feedback mechanism at the syllable level is appropriate. Finally, since inflection changes depending on the phrase type and a word's position in the phrase and, in some languages, the desired meaning, some sort of feedback must also exist at the word level, showing changes and transitions throughout the duration of each word and during the transition from one word to the next.

3. Error Portrayal

a. Wrong Phoneme

Identification of an error in pronunciation leads to the problem of how to present this error in a manner that is comprehensible by the student. The International Phonetic Alphabet contains symbols to portray each of the identified phonemes of all the world's languages. Table 2-2 shows these symbols only for the English language. While these symbols provide a meaningful, accurate, and unique way to represent each phoneme, they are typically unknown to a common student learning a new language. A more familiar method of representing pronunciation errors of phonemes needs to be used.

One such possibility is to present the proper pronunciation in a fashion similar to that found in a dictionary of the student's native language. While dictionaries also use special symbology, these symbols would be more familiar and a student would be able to consult their native language's dictionary for examples of these sounds in words in their native tongue. Consider the phrase "You like learning geography and chemistry." The International Phonetic Alphabet transcription of this word would be "ju lɑIk lɜrnIŋ ʤiɑgrəfi ənd kɛmIstri." A dictionary's transcription as found in the American Heritage Dictionary would be "yōō līk lûr'nĭng jē-ŏg'rə-fē ənd kĕm' ĭ-strē." A comparison of the two clearly shows that the latter is much easier to comprehend. A color scheme could be then used to signify whether a student's attempt at pronunciation is correct or incorrect. After an attempt at pronouncing the desired phrase, the dictionary pronunciation is then displayed. Black letters indicate satisfactory pronunciation and red letters represent improper pronunciation. Blue letters represent phonemes added by the student that are not present in the original. Yellow letters represent missing phonemes. Figure 3-2 shows a rendering of what this portion of the interface might look like.



You like learning geography and chemistry

yōō līk lûr'nĭng jē-ŏg'rə-fē ənd kĕm' ĭ-strē

Figure 3-2 Pronunciation Feedback for Phonemes[1]

[1] As a consequence of being printed in black and white, the color feedback suggested by this thesis can not be represented. In the actual interface, these errors are represented by the different colors indicated in the thesis.

b. Wrong Stress

The stress that is placed on a specific syllable is characterized by its respective intensity or loudness as compared to the other syllables. This requires that the feedback presented to a student be somehow displayed syllable by syllable with each syllable's intensity shown in respect to each of the other syllables. A stressed syllable corresponds to a louder or higher intensity syllable and an unstressed syllable is roughly of equal loudness or intensity of the other non-stressed syllables. The best way to provide this feedback to a student would be on a bar graph as shown in Figure 3-3. A line chart would not be as informative because the feedback needed to represent stress needs to represent relative amounts of stress in relation to other syllables. This relation between stress levels can be further exaggerated and emphasized through use of a bar graph with larger steps between adjacent syllables. Additionally the choice of a bar graph further illustrates that stress is essentially constant for a single syllable.

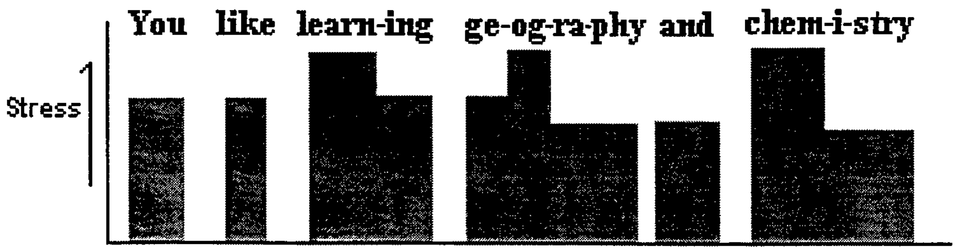


Figure 3-3 Stress Feedback

c. *Wrong Inflection*

Changes in inflection or tone are characterized by a change in the pitch of the voice of the speaker. An error in inflection can thus be defined as a change in the wrong direction or no change when a change is expected. This best way to visually present this information to the student would be in a similar fashion as stress except now the relative height represents pitch and not loudness. Although a bar graph could also be used to represent inflection, I chose a line graph for two reasons. First, this distinguishes the inflection feedback from the stress feedback by different visual presentations (live vs. bar graph). Second, it is possible and common for inflection to change across the duration of a phoneme, especially at boundaries. This requires a representation that changes quicker but still provides a smooth transition. Figure 3-4 shows how pitch information can be displayed.

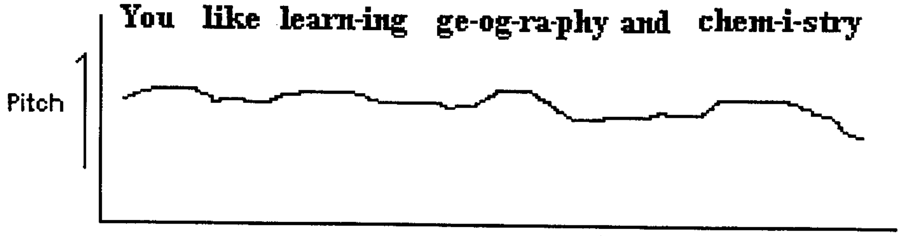


Figure 3-4 Pitch Feedback

B. COMBINING COMPONENTS INTO AN INTEGRATED INTERFACE

Putting all of the components together for an integrated product that provides all of the previously discussed feedback requires a few decisions as to the format in which the student will make pronunciation attempts. What kind of prompt the student will be given and which components will be displayed need to be addressed.

1. Student Input

I selected two modes of input, test and hint mode. These modes would allow the student to attempt pronunciation depending on their level of comfort or level of proficiency. Additionally, the test mode may be used to provide an unbiased assessment and possibly generate a score when testing students for pronunciation proficiency.

a. Test Mode

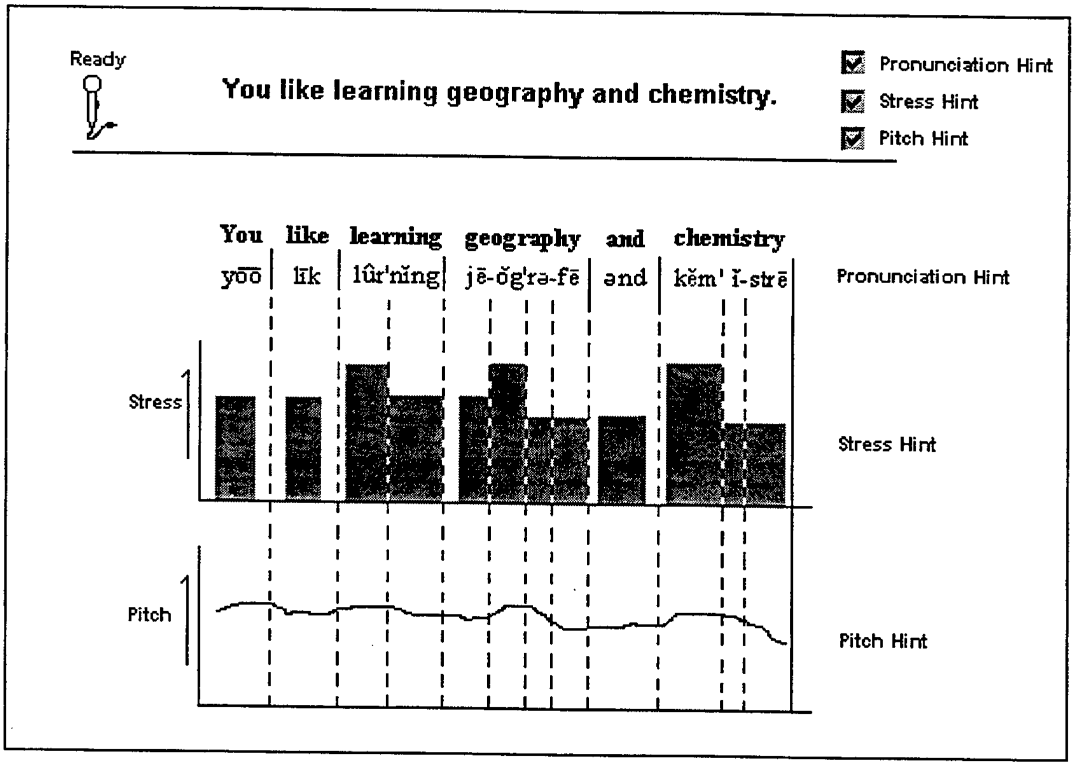
The test mode of operation would display only the phrase to be attempted to pronounce and a microphone indicator. The microphone would be red when the microphone is off and green when the microphone is turned on. This would then signal the student that they should make an attempt at pronouncing the displayed phrase. After the pronunciation attempt, feedback would then presented.

b. Hint Mode

The hint mode of operation would display the phrase to be attempted as well as the same red and green light indicator as the normal mode. However, the student would now be able to select any combination of hints, giving a visual guide to help the student in the pronunciation. Additionally, the phrase could be played for the student. The student could then practice the pronunciation several times using the information provided from the hints, concentrating on the area that they were unsure or weakest at pronouncing. Hints would be available in following areas:

- The phonetic dictionary-style pronunciation
- The desired stress graph
- The desired pitch graph

Figure 3-5 shows what all three the hints would look like.



You like learning geography and chemistry.

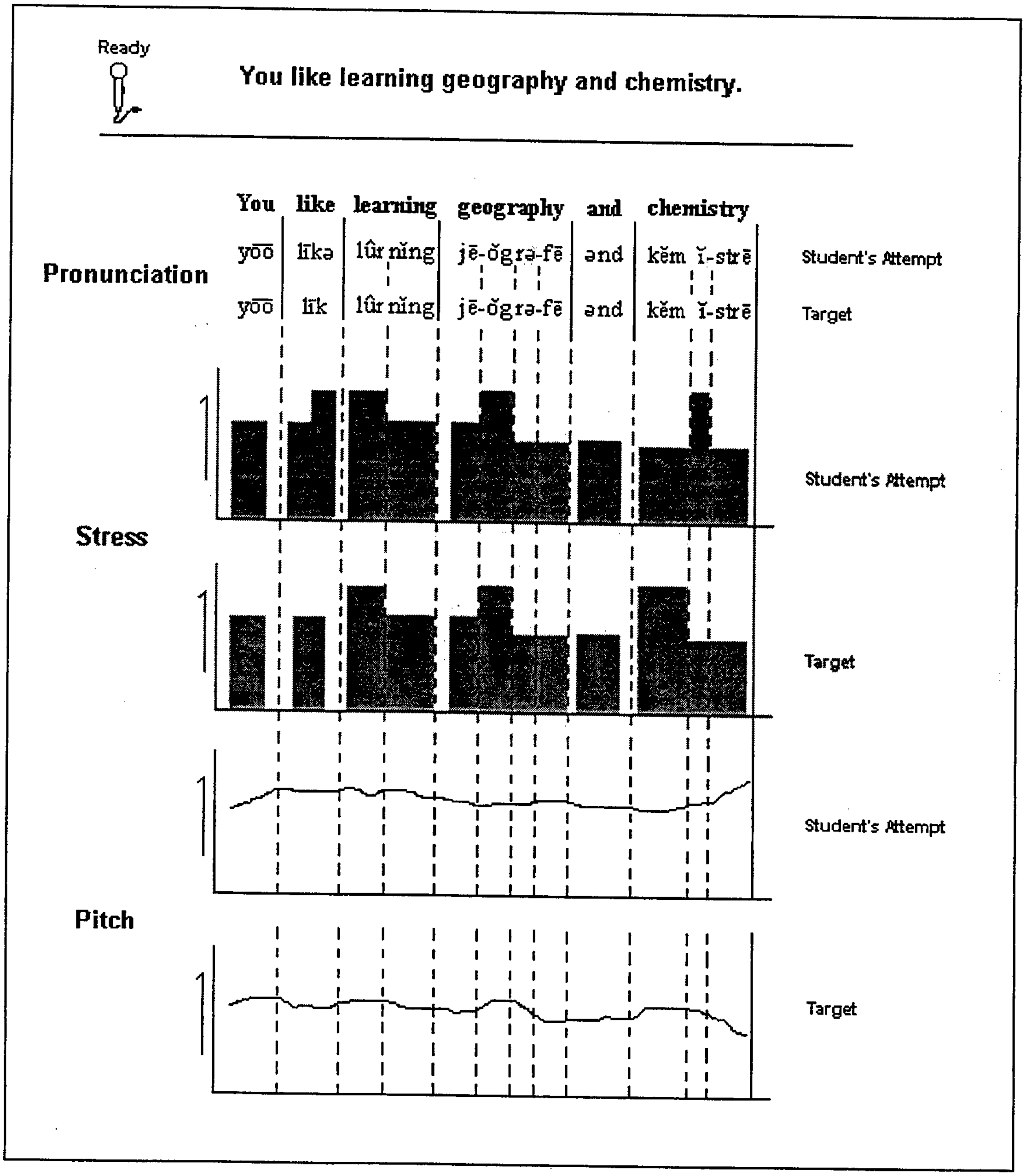
☑ Pronunciation Hint
☑ Stress Hint
☑ Pitch Hint

Ready
You like learning geography and chemistry
yōō līk lûr'nĭng jē-ŏg'rə-fē ənd kĕm' ĭ-strē
Pronunciation Hint
Stress
Stress Hint
Pitch
Pitch Hint

Figure 3-5 Pronunciation Hints

2. Feedback

In order to be effective, the feedback presented to the student would have to display the student's attempt at pronunciation and the desired or "ideal" pronunciation. There would be two methods of displaying the information. The first method would have each area of feedback for the student's attempt, (phoneme, stress and pitch), stacked directly above the same portion of the desired "target". Each area would then be stacked above each other so that corresponding portions are in alignment. This method of stacking allows the student to see how their attempt lines up with the target and how the stress and pitch change with each portion of pronunciation. Figure 3-6 depicts this method.



Ready

You like learning geography and chemistry.

Figure 3-6 Feedback by stacking Student's Pronunciation above the Target[2]

[2] The colors of the interface can not be depicted adequately in black and white. In color the student's pronunciation attempt contains an example of the colors used to represent the incorrect phonemes. The word *like* has an extra phoneme represented in blue, the word *geography* has an incorrect phoneme represented in red and the student did not pronounce the second syllable in the word *chemistry*, indicated in yellow.

An alternate method of displaying the feedback would superimpose the student's attempt directly on top of the target pronunciation (Figure 3-7). This would allow the student to more accurately compare the relative height between their attempt and the target than the first method. However it makes visualization a little more difficult in areas where the graphs are equal but it highlights the differences better.

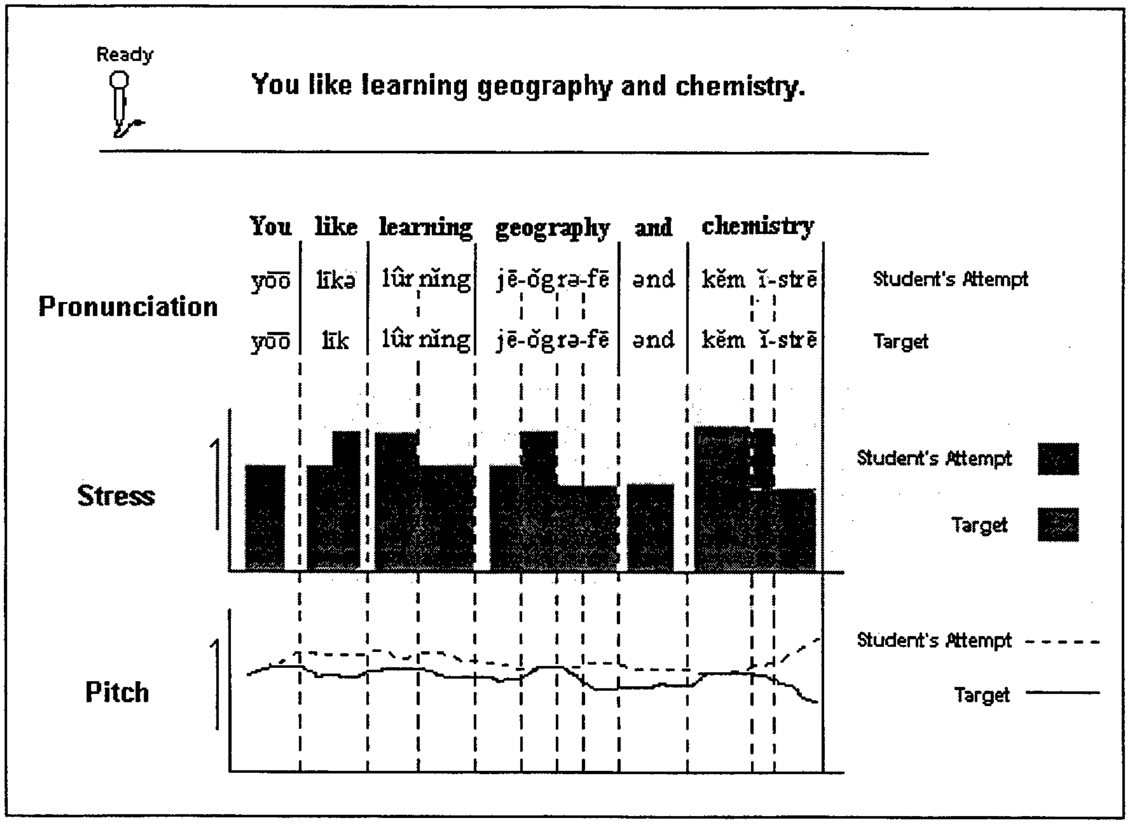


Figure 3-7 Feedback by Superimposing Student's Pronunciation on top of Target[3]

[3] Again the colors can not be seen. The student's pronunciation attempt has the same colors as described in the footnote for Figure 3-6. However, the stress bar graph now has regions of contrasting colors to portray the regions where the student's attempt differs from the target.

C. SUMMARY

This chapter discussed how the components were chosen in designing the interface to provide feedback for a student learning language pronunciation. The types of errors, how these errors could be presented and how these forms of feedback could be presented in an integrated, configurable interface were all discussed.

IV. SYSTEM DESIGN

A. INTRODUCTION

In order to provide the interface with necessary data, it becomes necessary to do some sort of digital signal processing similar to conventional speech recognition. However, this is not speech recognition because speech recognition analyzes a digital speech signal for certain key characteristics and then uses these characteristics to find the closest match from a word database. In providing pronunciation feedback to a student, this type of system would be ineffective because if the student's pronunciation perfectly matches the database for 80% of the word, that word could be identified as the speech produced and there is no reference to identify the 20% of the word that is in error. In order to be able to provide feedback, the system must identify which specific areas are incorrect, in segments smaller than words.

B. PROPOSED METHODOLOGY

Since the smallest unit of required feedback is the phoneme, a naïve approach to this problem would be to have an instructor speak into a phoneme dictation device and have a student speak into this same device. The results would be compared and feedback displayed. Research to develop this "phonetic typewriter" has been ongoing since the 1960's and "much more is now known about the complex interactions between the parts of the acoustic signal that make it impossible to apply any such sequential recognition strategy" [Ref. 1]. An alternative approach needs to be considered.

1. Generating the Ideal or "Target" Signal

The first step in building the "target" signal would be to record the desired phrases multiple times from multiple native speakers. Since there is a significant difference, particularly in pitch, between males and females, these recorded speech segments should be divided based on sex. For each sex group, the signals should then be normalized, that is dividing each sample in the signal by the maximum amplitude present in that signal. This results in all samples in each signal having an amplitude in the range from zero to one. Next some sort phoneme boundary identification needs to be accomplished either through automated means or manually through listening. Once the phoneme boundaries are identified, the signals can then be compared phoneme by phoneme.

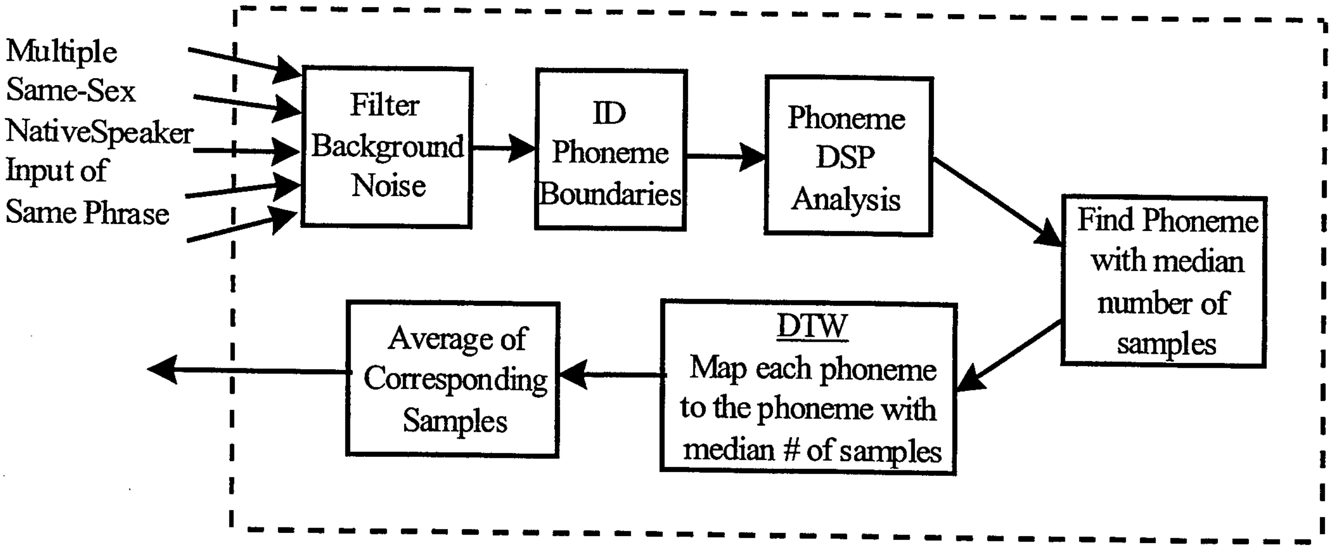

Multiple
Same-Sex
NativeSpeaker
Input of
Same Phrase
Filter Background Noise
ID Phoneme Boundaries
Phoneme DSP Analysis
Find Phoneme with median number of samples
DTW
Map each phoneme to the phoneme with median # of samples
Average of Corresponding Samples

Figure 4-1 Building the "Target" Signal

Once each phoneme is identified, some sort of digital signal processing must be done to identify and extract the key characteristics necessary for later analysis and comparison. Since it is very unlikely that two utterances of the same word will ever be of the same length in time, phonemes will be of different duration, represented digitally by different numbers of samples for each phoneme. More analysis is needed to align phonemes of different time length.

To line up corresponding samples a process know as Dynamic Time Warping (DTW) can be used. What DTW does is using the shortest path algorithm, it compares characteristics of each sample and develops a mapping of one set of samples to another. For each phoneme in the phrase, a phoneme signal segment with the median number of samples should be selected. All of the other corresponding phoneme segments from the other signals should be mapped to it using DTW. By selecting the phoneme with the median number of samples results in half of the signals having one to many mappings and half having many to one mappings between it and the signal with the median number of samples. Once this mapping is complete, a new "average" signal can be generated based on averaging the characteristics of corresponding samples. This can then be used as the "target" signal to compare the student's attempt against.

2. Comparing the "Target" Signal to the Student's

Once a target signal has been developed, it can now be used as the reference signal with which to compare the student's pronunciation against and from which to develop feedback. The overall concept of developing feedback is depicted in Figure 4-2. The middle two boxes are the points of discussion in the paragraphs that follow.

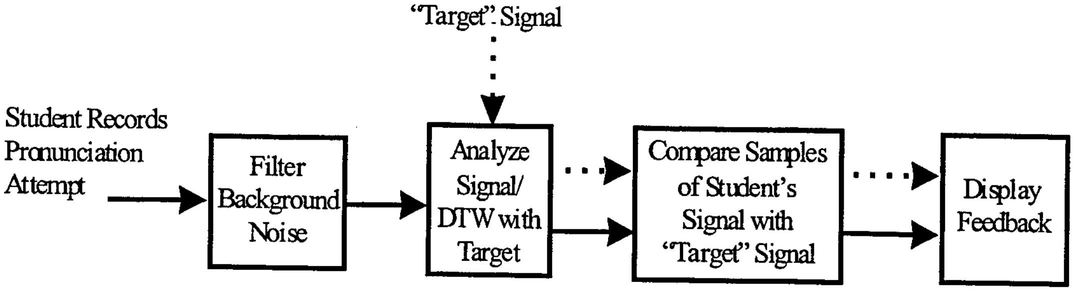

"Target" Signal
Student Records Pronunciation Attempt
Filter Background Noise
Analyze Signal/ DTW with Target
Compare Samples of Student's Signal with "Target" Signal
Display Feedback

Figure 4-2 Feedback Development

After filtering out background noise, the student's digital signal needs to go through the same series of digital signal processing steps to extract the same sort of key characteristics that were extracted for the "target" signal. Using these characteristics, the DTW procedure can then be applied to map samples of the student's signal to the "target" signal. Once the samples are mapped, the phoneme boundaries for the student's signal can be inferred from the "target" signal. These boundaries can then be used to compare phonemes and to determine if the student's phonemes are the same as the target's. When forming the "target" signal, a statistical analysis can be conducted to determine the standard deviation of all the pieces of the "target" signal. If the student's phoneme is within a standard deviation of the "target" then the feedback would indicate it as acceptable (black letters). If it is outside a standard deviation, then the letters would be red, indicating an improper phoneme.

Since in the preceding analysis the basic unit is the phoneme, it makes sense that while using DSP techniques for phoneme analysis that at the same time for the same sampling windows, that pitch and stress data be determined. For each phoneme this information is stored and used for feedback. For pitch, the resulting graph would require some smoothing to connect the discontinuities resulting from unvoiced segments and moments of silence.

C. LISP IMPLEMENTATION

LISP is a relatively simple programming language that is easily extensible and allows a programmer to write sophisticated programs faster and in less lines of code. As such it has become a popular language for rapid prototyping. This allows a programmer to quickly implement a working prototype that can be later used as the spec for implementing the system in any other high level language [Ref. 5]. In the sections that follow, a LISP implementation of certain objects required for the system described by this thesis is discussed.

1. The Signal Class

The signal object represents any digital signal. As such, it contains data slots for information that describes the different characteristics and components of a digital signal. The *name* data slot provides a way to uniquely identify different signals. The *file-name* slot is used to store the file name of the associated digital signal. The *number-of-samples* slot stores the number of samples present in the digital signal and the *time-length* slot describes the duration of the signal down to milli-seconds. The slot labeled *maximum* represents the highest amplitude present in the signal. The *spectrogram* and *freq-vs-t-graph* slots are used to store the names of the associated graphic files that show these graphs for this signal. The *number-of-phone* slot is the number of phonemes contained in the signal. *Phone-list* is a list of all the phoneme objects present in the signal. Once instantiated, a signal class object can be manipulated to display information required for the graphical output of the interface.

```
(defclass signal-class ()
  ((name
    :initform 'no-name
    :accessor name)
   (file-name
    :initform ‘this-file
    :accessor file-name)
   (number-of-samples
    :initform 101
    :accessor number-of-samples)
   (time-length
    :initform 0
    :accessor time-length)
   (maximum
    :initform 0
       :accessor maximum)
   (spectrogram
    :accessor spectrogram
    :initform "sample-sg")
   (freq-vs-t-graph
    :accessor freq-vs-t-graph
    :initform "sample-fg")
   (number-of-phone
    :initform 0
    :accessor number-of-phone)
   (phone-list
    :accessor phone-list
    :initform '())))
```

Figure 4-3 Signal Class Lisp Code

2. The Phoneme Class

As described in previous sections of this chapter, the smallest unit of analysis is the phoneme. As such, phonemes could be modeled in LISP as a unique Phoneme Class. However, phonemes are actually segments of the original signal or sub-signals. As such, they need to include all the same information as signal objects plus some additional information. This is accomplished through inheritance and explains why the phone-class object is defined with the signal-class immediately after in parenthesis. The code segment below shows that all of the information required to identify the phoneme and represent the phoneme in the interface is present. Once the results of the digital signal processing is accomplished, the data required for each phoneme object can then be assigned into the proper slot. This information can then be later used to supply the interface with data for displaying the entire signal.

```
(defclass phone-class (signal-class)
   ((position-of-phone
     :initform 1
     :accessor position-of-phone)
    (start-time
     :initform 0
     :accessor start-time)
    (end-time
     :initform 0
     :accessor end-time)
    (start-pitch
     :initform 0
     :accessor start-pitch)
    (end-pitch
     :initform 0
     :accessor end-pitch)
    (stress
     :initform 0
     :accessor stress)
    (dictionary-representation
     :initform h
     :accessor dictionary-representation)))
```

Figure 4-4 Phoneme Class Lisp Code

The *position-of-phone* slot refers to where this phoneme appears in the original signal relative to all the other phonemes. The *start-time* and *end-time* slots refer to the phoneme boundary of this phoneme in the original signal in seconds. Seconds were chosen over samples for no other reason then to give accessibility to this data member in units that are meaningful to a programmer. The *start-pitch* and *end-pitch* slots represent the pitch at the start and end of the phoneme. This is significant because it is possible for the pitch to gradually change across a phoneme particularly at the boundaries where coarticulation is occurring. Since in the interface pitch is being displayed as a line graph, this helps have a gradual change in the line representing the pitch as opposed to a potential step effect similar to a bar graph if pitch was only measured and stored once for each phoneme. The *stress* slot represents the average stress or intensity of the signal in the phoneme and the *dictionary-representation* slots represents the string that phonetically represents the phoneme as would be found in the dictionary.

3. Pitch Display

In the implementation of phoneme and signal objects, numerical data necessary for displaying pitch and stress is located in the phoneme object. The only access to the phoneme objects is through the signal object which has a data slot containing a list of phoneme objects that comprise the signal. To display the pitch contour or the stress bar graph of a signal, the signal's phoneme list is used to step through all of the phoneme objects to access and display the pitch and stress information for each phoneme.

A Lisp implementation of stepping through each of the phonemes in a signal object is shown in Figure 4-5. To display stress information, a similar code segment would be used but instead of having a start and end stress, a constant value for stress is used for start and end stress for each phoneme. The last line of the code executes a method called *graph-phone-pitch* for each phoneme on the signal's *phone-list* and actually draws the line graph. The other lines of the code make, initialize, size and name a window object and then draws the axis in the center of the window. The second to last line assigns the amount of leading blank space (or silence) before the first phoneme in the signal to a variable named *leading-x-space.*

```
((defmethod graph-signal-pitch ((signal signal-class))
   (setf pitch-win
         (make-window-stream :borders 1
                             :left 300
                             :bottom 300
                             :width 500
                             :height 300
                             :title (format nil "~A Signal's Average
                                                 Pitch"(name signal))
                             :activate-p t))
   (draw-axis)
   (setf leading-x-space (start-time (car (phone-list signal))))
   (mapcar 'graph-phone-pitch (phone-list signal)))
```

Figure 4-5 Lisp Code to Graph a Signal's Pitch

The *graph-phone-pitch* is shown in Figure 4-6. The first line of code sets a variable named *duration* equal to the phoneme's duration. This is simply the difference between the phoneme data slots *start-time* and *end-time* of the phoneme multiplied by a scaling factor **x-scale**. This scaling factor represents the number of horizontal pixels used to represent one second in time along the x-axis. The third line initializes the *xpos* variable to shift the pitch contour to the left to compensate for and disregard leading silence. Then the fourth and fifth line position the window cursor to the proper position to write the name of the phoneme accomplished with the sixth line of the code. The three lines in the next segment of the code draw a vertical line to separate phonemes. The final portion of the code actually positions the location for and draws a single pixel for each value specified by the *duration* variable set in the first line in the code. It is assumed that pitch changes uniformly across the duration of the phoneme from the *start-pitch* to the *end-pitch* values used to represent the measured pitch at the start and end of the phoneme. A **y-scale** variable is used like the *x-scale* variable to represent the number of pixels used to represent a change in pitch of one hertz.

```
(defmethod graph-phone-pitch ((phone phone-class))
  (let ((duration (* *x-scale* (- (end-time phone) (start-time phone))))
        (xpos (* *x-scale* (- (start-time phone) leading-x-space))))
    (setf (window-stream-y-position pitch-win) 100)
    (setf (window-stream-x-position pitch-win) xpos)
    (format pitch-win "~A" (name phone))  ;; write phoneme name

    ;; Make a vertical line to separate phones
    (dotimes (i 100)
      (draw-point-xy pitch-win xpos (+ i 108)))

    ;; draw pitch contour for the duration of the phone
    (dotimes (i duration)
      (draw-point-xy pitch-win (+ xpos i)   ;;x coord
                               (+ (* *y-scale* (+ (start-pitch phone)
                                                (* (- i 1)
                                                   (/ (- (end-pitch phone)
                                                         (start-pitch phone))
                                                   duration))))
                                  *y-origin*)))  ;; y-coord))
```

Figure 4-6 Lisp Code to Graph a Phoneme's Pitch Contour

The resulting display for the LISP code presented in Figures 4-5 and 4-6 is shown in Figure 4-7.

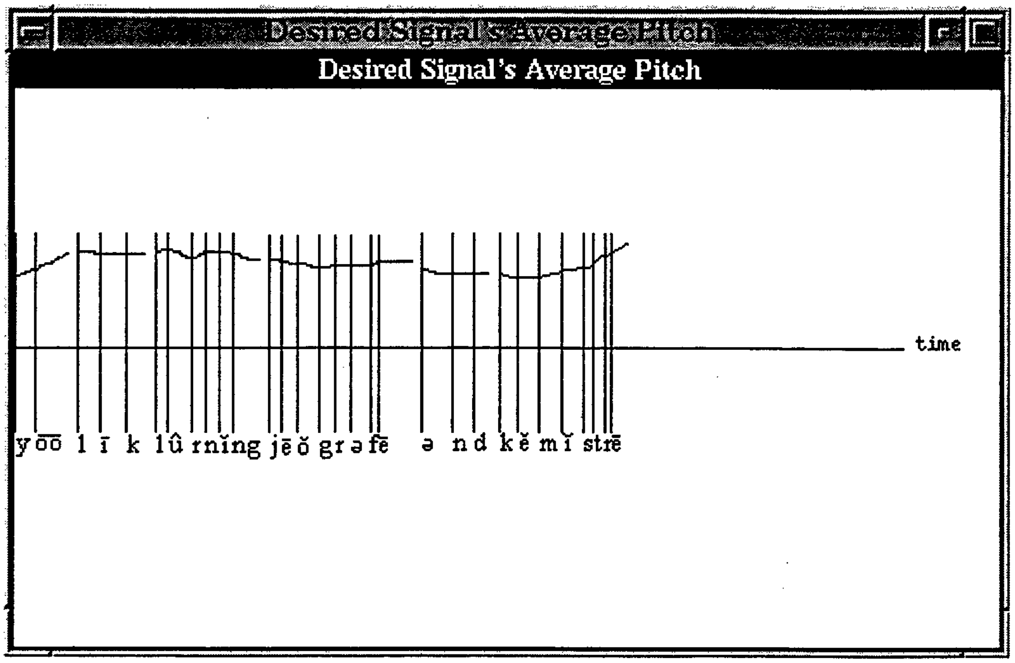

Desired Signal's Average Pitch
Desired Signal's Average Pitch
time
yo͞o l ī k lû rnĭng jē ŏ grə fē ə nd kĕ mĭ strē

Figure 4-7 Lisp Code Output for Pitch

D. SUMMARY

This chapter starts by proposing a possible technique for making a "target" signal from multiple recorded digital signals from multiple speakers. This technique required that the recorded signals are filtered, normalized and phoneme boundaries identified. The target signal was then built phoneme by phoneme by taking the phoneme with the median number of samples and mapping it's samples to the samples of the same phoneme in the other signals using a technique called Dynamic Time Warping (DTW). After DTW, the signals could be "averaged" and statistical data formed. Then a students attempt at pronunciation would be recorded. The resulting signal then undergoes DTW with the target signal and phoneme boundaries are inferred. Each phoneme can then be analyzed individually and compared. The resulting data from the comparison is then supplied to the interface described in Chapter III. The chapter concludes with a LISP prototype that shows how the signals and phonemes could be modeled and data extracted to form the required graphs.

V. CONCLUSION

A. SUMMARY

This thesis offers a design for an interface to provide feedback for a student learning language pronunciation. Students learning pronunciation make errors in three general areas, namely, improper pronunciation of the individual speech sound or phoneme, improper application of stress on the wrong part of a word, and improper intonation on the wrong part of the word or phrase. Each of these error areas require their own indicator for visual feedback. Each must specify the part of the word or phrase that is incorrect and what kind of error was made. Additionally, the desired or "target" pronunciation must be provided along with the student's attempt to allow the student to see the error and understand the type and magnitude of the error (i.e. not enough stress, tone too high, etc.). Stacking or superimposing the three forms of feedback allows the student to line up like segments and rapidly identify which segments are in error in which areas. Additionally, giving the student two modes of operation allows the student the opportunity to receive help prior to their pronunciation attempt or to test themselves by making an attempt with no hints from the system.

The digital speech processing to supply this interface with data remains to be developed. Current speech processing theory and technology just now may be able to accurately build a representative target pronunciation signal, make the necessary comparisons, and supply the necessary data for feedback. One such suggestion for this type of system is the focus of Chapter IV. The interface components and design are ready - all that is needed is the speech processing to provide the data.

B. READDRESS THESIS QUESTIONS

This goal of this thesis was to examine the requirements of and create an interface that supplies useful information to a student learning pronunciation of a language. Specifically the following research questions were addressed:

- How is speech produced and how can it be classified and represented?
- What causes pronunciation errors and in what areas might a student attempting to learn proper pronunciation make errors?
- What areas must an interface address to provide useful visual feedback to a student attempting to learn a foreign language?
- Can this interface adequately integrate and represent pronunciation errors?
- What is a possible approach using speech processing (digital signal processing) techniques to identify errors and supply the interface useful data?

How is speech produced and how can it be classified and represented? Chapter II described how speech production starts as air is forced from the lungs through the vocal folds in the larynx. Based upon the distance between the chords, the resulting sound can be voiced, aspirated or unvoiced. Voiced speech segments are vowels, unvoiced, consonants. The air continues through the various parts of the vocal tract, with each part having an opportunity to alter or change the sound. Consonants are classified based upon the place of articulation and the manner of articulation. The possible consonant classifications based on place of articulation are labial, labiodental, dental, alveolar, palatal, velar, and glottal. The possible consonant classifications based on the manner of articulation include plosives, nasals, fricatives, affricatives and approximants. Vowels are classified based on tongue height at the place of greatest closure and the part of the mouth of the horizontal place of greatest point of closure. High, middle and low are used to describe the former; front, central and back, the latter. Vowels are also described based on whether the lips are rounded or not. Finally, a special classification of vowels exists known as the diphthongs. Diphthongs are vowels that begin with the mouth in the position of one vowel and transitions and ends with the mouth in the position of another. Speech can be recorded and represented as a digital signal, which depends on a specific sampling rate and the number of quantizization levels. The number of quantizization levels determines the number of bits required for one sample and the sampling rate determines the number of samples required for one second of speech. The smallest segment of pronunciation is the phoneme. The International Phonetic Alphabet contains unique symbols to represent all of the possible speech sounds in all the world's languages. Being awkward to use and understand, ARPA developed two alphabets (ARPAbets) that use standard English letters, a single letter and, a two letter uppercase alphabet.

What causes pronunciation errors and in what areas might a student attempting to learn proper pronunciation make errors? In Chapter III, I stated that errors in pronunciation can be attributed to three general areas: the student improperly forming the different parts of the vocal tract and producing the wrong speech sounds; the student's improper application of stress to the wrong syllable; the student's improper change in tone or pitch of segments of the word or phrase. Production of the wrong speech sounds can be attributed to the student configuring the different parts of the vocal tract improperly, either due to lack of familiarity or inability. Stress and pitch errors can be attributed to lack of familiarity.

What areas must an interface address to provide useful visual feedback to a student attempting to learn a foreign language? Also in Chapter III, I proposed that at minimum, an interface must provide information to the student concerning speech sounds (phonemes), pitch, and stress. In order to be of use, the interface must provide clues as to what specific segment is in error and what the error specifically is, and perhaps how to fix the error. For phoneme production, this information must identify the part in error, point out what was said and point out what should have been said. Errors need to be highlighted in some fashion, (e.g. red letters for improper phonemes and blue letters for additional phonemes not included in the original word or phrase). For stress, the interface must provide information as to what the desired areas of stress are and what areas the student actually stressed. This information should be provided at the syllable level since this is the basic unit at which stress is applied. A bar graph serves this purpose well. Finally, for pitch, the interface needs to provide the same sort of information as provided for stress. However, since pitch can change across a phoneme, particularly at phoneme boundaries, pitch needs to be displayed in a manner that provides for rapid change over a small interval. For this display, a line graph serves well.

Can this interface adequately integrate and represent pronunciation errors? In order to be effective, the interface needs to be able to display all of the information on the errors in a format so that the student can see the error in each individual area. Chapter III concludes by describing this integration of displaying the errors, particularly stacking. Stacking each individual feedback area on top of each other with the individual word segments (phoneme and/or syllables) lined up gives the student an opportunity to quickly see which parts of the phrase are incorrect. In addition to providing feedback on correctness, the interface provides some sort of relative measure as to how close or how far from the desired pronunciation the attempt was. Displays of the student's attempt and the desired pronunciation are either stacked on top of each other or superimposed one on top of the other. Both have similar segments lined up and the superimposed portrayal would give the student an opportunity to make a more accurate assessment on the relative closeness of their attempt to the desired pronunciation.

What is a possible approach using speech processing (digital signal processing) techniques to identify errors and supply the interface useful data? Chapter IV described that the first thing that would need to be done is to develop a target signal with which to compare the student's signal against. Multiple recordings would need to be averaged into one signal with statistical analysis defining a set of acceptable limits. In order to do this, the signals need to be "prepared." Different speech signals vary in relative amplitude and duration. To have the signals comparable in regard to amplitude, a technique known as normalization is used. For each signal the maximum amplitude is determined and this value is then used to divide each amplitude of all of the samples in the signal. This results in each amplitude in the signal being between a value of zero and one. Next, a way to line up corresponding samples needs to be examined. For each signal, phoneme boundaries would need to be identified. Corresponding speech signals will contain a different number of samples and a way is needed to match corresponding samples, especially since in some case one sample in signal A will correspond to two samples in signal B, or visa versa. Dynamic Time Warping (DTW) is the technique used to map corresponding samples of different signals. Once the samples are matched, a "target" signal can be constructed. When the student makes an attempt at pronunciation, the signal is recorded, normalized and undergoes DTW to map its samples to the target signal. The results of the DTW allow the phoneme boundaries to be inferred onto the student's signal from the target signal. Then a comparison of the student's and target signal can be conducted for each phoneme. Comparison of the pronunciation, pitch and stress can be conducted and the results graphically displayed on the interface.

C. RECOMMENDATIONS FOR FUTURE RESEARCH

Future research in this area should concentrate on collecting the digital signal of multiple native speakers and constructing a "target" signal as described in this thesis. To see if the proposed system would work, five to ten recorded speech signals could be used from as little as two or three speakers to test the concept. It would be interesting but not necessary to see if the resulting target signal can be recognized as intelligible speech.

Once the target signal is obtained, recording and inferring phoneme boundaries on a student's attempt should be then be tried. Once inferred, each segment should be played back individually to verify accuracy of phoneme boundary identification. If this proposal of using DTW to infer phoneme boundaries is imperfect, perhaps this method can be used to obtain a quick general area of the boundary which can then further be refined using other signal information such as zero crossing rate, short term energy, or presence or absence of certain frequencies.

Once this process of arriving at phoneme boundaries in the student's attempt is successful, similar phonemes should then be analyzed and compared and plotted on the interface denoting feedback in pronunciation, stress and pitch. Once a working interface is developed, further exploration into configurable options such as colors, size, and 3D bar graphs should be researched and feedback on their effectiveness solicited from foreign language students and instructors alike.

This thesis makes a good first step at determining what type of feedback a student actually requires. Most of the future work required is in the actual digital signal processing, dynamic time warping, word and phoneme boundary identification and phoneme recognition. Clearly, a student requires feedback to learn pronunciation. Visual feedback, as well as acoustical, can help a student to better understand their errors and this thesis offers such an approach.

APPENDIX A: SOURCE CODE (SIGNAL CLASS)

```
;;-------------------------------------------------------------
;; File: signal.lsp              Franz Common Lisp
;; Ken Fritzsche
;; 21 August 97
;; Contains the Signal class
;; Copyright © 1997 Ken Fritzsche
;;-------------------------------------------------------------

        (require :xcw)
        (use-package :cw)
        (initialize-common-windows)

        (defclass signal-class ()
         ((name
           :initform 'no-name
           :accessor name)
          (number-of-samples
           :initform 101
           :accessor number-of-samples)
          (time-length
           :initform 0
           :accessor time-length)
          (bandwidth
           :initform '(0 0)
           :accessor bandwidth)
          (maximum
           :initform 0
           :accessor maximum)
          (spectrogram
           :accessor spectrogram
           :initform "sample-sg")
          (freq-vs-t-graph
           :accessor freq-vs-t-graph
           :initform "sample-fg")
          (number-of-phone
           :initform 0
           :accessor number-of-phone)
```

```
   (phone-list
    :accessor phone-list
    :initform '())))


(defmethod initialize-signal ((signal signal-class) signal-data)
  (setf (name signal) (first signal-data))
  (setf (number-of-samples signal) (second signal-data))
  (setf (time-length signal) (third signal-data))
  (setf (bandwidth signal) (fourth signal-data))
  (setf (maximum signal) (fifth signal-data))
  (setf (spectrogram signal) (sixth signal-data))
  (setf (freq-vs-t-graph signal) (seventh signal-data))
  (setf (number-of-phone signal) (eighth signal-data))
  (initialize-phone-list signal (ninth signal-data)))


(defmethod initialize-phone-list ((signal signal-class) phone-name-list)
  (do* ((counter (- (length phone-name-list) 1) (- counter 1)))
      ((< counter 0) 'done)
      (setf (phone-list signal) (cons (make-instance 'phone-class) (phone-list signal)))
      (initialize-phone (car (phone-list signal)) (nth counter phone-name-list) counter)))


(defmethod initialize-phone ((phone phone-class) phone-data-list posit)
  (setf (name phone) (first phone-data-list))
  (setf (position-of-phone phone) posit)
  (setf (start-time phone) (second phone-data-list))
  (setf (end-time phone) (third  phone-data-list))
  (setf (start-pitch phone) (fourth phone-data-list))
  (setf (end-pitch phone) (fifth phone-data-list))
  (setf (freq-vs-t-graph phone) (sixth phone-data-list)))


(defmethod show-freq-graph (pathname)
  (setf win2
    (make-window-stream :left 0 :width 700
                  :bottom 100 :height 300
                  :activate-p t : title "Amplitude vs. Time"))
  (bitblt (read-bitmap pathname : format :x11) 0 0 win2 0 0))
```

```
(defun show-spectrogram (pathname)
  (setf win3
    (make-window-stream :left 0 :width 500
                      :bottom 100 :height 500
                      :activate-p t : title "Spectrogram"))
   (bitblt (read-bitmap pathname : format :x11) 0 0 win3 0 0))


(defmethod print-phones ((signal signal-class))
  (format t "The signal ~A contains the following phones: ~%" (name signal))
  (do ((the-phone-list (phone-list signal) (cdr the-phone-list)))
     ((null the-phone-list) 'done)
     (format t "~A~%" (name (car the-phone-list)))))
```

APPENDIX B: SOURCE CODE (PHONEME CLASS)

```
;;----------------------------------------------------------------
;; File: phone.lsp              Franz Common Lisp
;; Ken Fritzsche
;; 21 August 97
;; Contains the Phoneme class
;; Copyright © 1997 Ken Fritzsche
;;----------------------------------------------------------------

        (defclass phone-class (signal-class)
         ((position-of-phone
           :initform 1
           :accessor position-of-phone)
          (start-time
           :initform 0
           :accessor start-time)
          (end-time
           :initform 0
           :accessor end-time)
          (start-pitch
           :initform 5
           :accessor start-pitch)
          (end-pitch
           :initform 9
           :accessor end-pitch)))
```

APPENDIX C: SOURCE CODE (PITCH GRAPH)

```
;;-----------------------------------------------------------------
;; File: graph-pitch.lsp                Franz Common Lisp
;; Ken Fritzsche
;; 21 August 97
;; Contains the methods to print the pitch of a signal class object
;; Copyright © 1997 Ken Fritzsche
;;-----------------------------------------------------------------

      (require :xcw)
      (use-package :cw)
      (initialize-common-windows)

      (setf *y-origin* 150)
      (setf *y-scale* 10)
      (setf *x-scale* 200)


      (defmethod graph-signal-pitch ((signal signal-class))
        (setf pitch-win (make-window-stream :borders 1
                                 :left 300
                                 :bottom 300
                                 :width 500
                                 :height 300
                                 :title (format nil "~A Signal's Average Pitch"(name signal))
                                 :activate-p t))
        (draw-axis)
        (setf leading-x-space (start-time (car (phone-list signal))))
        (mapcar 'graph-phone-pitch (phone-list signal)))


      (defmethod graph-phone-pitch ((phone phone-class))
       (let ((duration (* *x-scale* (- (end-time phone) (start-time phone))))
           (xpos (* *x-scale* (- (start-time phone) leading-x-space))))
         (setf (window-stream-y-position pitch-win) 100)
         (setf (window-stream-x-position pitch-win) xpos)
         (format pitch-win "~A" (name phone))
         ;; Make a vertical line to separate phones
         (dotimes (i 100)
          (draw-point-xy pitch-win xpos (+ i 108)))
         ;; draw actual pitch for the duration of the phone
```

```
   (dotimes (i duration)
    (draw-point-xy pitch-win (+ xpos i)   ;;x coord
                      (+ (* *y-scale* (+ (start-pitch phone)
                                  (* (- i 1)
                                    (/ (- (end-pitch phone)
                                      (start-pitch phone))
                                    duration))))
                        *y-origin*)))  ;; y-coord
))


(defun draw-axis ()
  (dotimes (i 450)
    (draw-point-xy pitch-win i *y-origin*))
    (setf (window-stream-y-position pitch-win) 150)
    (setf (window-stream-x-position pitch-win) 455)
    (format pitch-win "time ~%")
  (finish-output pitch-win)
)
```

LIST OF REFERENCES

1. Marcus, Stephen M., Syrdal, Ann K., "Speech: Articulatory, Linguistic, Acoustic, and Perceptual Descriptions," in *Applied Speech Technology,* edited by A. Syrdal, R. Bennett, and S. Greenspan, pp. 1-43, 1995.

2. O'Shaughnessy, Douglas, "Speech Technology," in *Applied Speech Technology,* edited by A. Syrdal, R. Bennett, and S. Greenspan, pp. 46-95, 1995.

3. Gopal, H.S., "Technical Issues Underlying the Development and Use of a Speech Research Laboratory," in *Applied Speech Technology,* edited by A. Syrdal, R. Bennett, and S. Greenspan, pp. 315-341, 1995.

4. Cantrell, M., *Speech Recognition Using Artificial Neural Networks*, Master's Thesis, Naval Postgraduate School, Monterey, California, March 1996.

5. Graham, P., "*ANSI Common Lisp*," Prentice Hall, Englewood Cliffs, New Jersey, 1996.

6. Neumeyer, Leonardo, Franco, Horacio, Weintraub, Mitchel, Price, Patti, "Automatic Text-Independent Pronunciation Scoring of Foreign Language Student Speech", Speech Technology and Research Laboratory, SRI International, 1996.

7. Advertisement for EZ Language Software, IMSI, 1997.

8. Advertisement for Syracuse Language Systems series of language software, 1997.

9. Advertisement for The Learning Company series of language software, 1997.

10. Parkes, Carl., *Thailand Handbook, Second Edition*, Moon Publishing, Chico, California, 1997.

BIBLIOGRAPHY

Anderson-Hsieh, Janet, "Interpreting Visual Feedback on Suprasegmentals in Computer Assisted Pronunciation Instruction", in *CALICO Journal,* Volume 11, Number 4, Summer 94.

Chen, S., and Chan, M., "Challenges in Robust Automatic Speech Recognition", presented at U.S. Army Research Lab Advanced Display & Interactive Display Consortium, January 1997.

Cooke, Martin, Beet, Steve, and Crawford, Malcom, *Visual Representations of Speech Signals,* John Wiley & Sons, New York, N.Y., 1993.

The Computer Assisted Language Instruction Consortium (CALICO) Home Page, http://calico.org/

Computer Assisted Pronunciation Investigation Teaching and Learning (CAPITAL) Special Interest Group Home Page, http://showme.missouri.edu/~langdans/capital.html

Cooke, Martin, Beet Steve and Crawford, Malcom., *Visual Representations of Speech Signals,* John Wiley & Sons, 1993.

Fidelman, Carolyn G., "A Language Professional's Guide to the World Wide Web", in *CALICO Journal,* Volume 13, Number 2 & 3, Winter 95/Spring 96.

Mammone, Richard J., Zhang, Xiaoyu, and Ramachandran, Ravi P., "Robust Speaker Recognition", in *IEEE Signal Processing Magazine,* September 1996.

Nagata, Noriko, "Computer vs. Workbook Instruction in Second Language Acquisition", in *CALICO Journal,* Volume 14, Number 1, Fall 96.

Proakis, John G. and Manolakis, Dimitris G., *Digital Signal Processing,* Third Edition, Prentice Hall, Upper Saddle River, New Jersey, 1996.

Spaai, Gerard W. G., and Hermes, Dik, "A Visual Display for Teaching Intonation", in *CALICO Journal,* Volume 10, Number 3, Spring 93.

Stenson, Nancy, Downing, Smith, Jan, and Smith Karen, "The Effectiveness of Computer Assisted Pronunciation Training", in *CALICO Journal,* Volume 9, Number 4, Summer 92.

Syrdal A., Bennett R., and Greenspan S., *Applied Speech Technology,* CRC Press, Boca Raton, Florida, 1995.

Tufte, Edward R., *Envisioning Information,* Graphics Press, Cheshire, Connecticut, 1990.

Young, Steve, “A Review of Large-vocabulary Continuous-speech Recognition”, in *IEEE Signal Processing Magazine*, September 1996.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center..2
8725 John J. Kingman Road, Ste 0944
Ft. Belvoir, VA 22060-6218

2. Dudley Knox Library 2
Naval Postgraduate School
411 Dyer Rd.
Monterey, CA 93943-5101

3. Chairman, Code CS 1
Computer Science Department
Naval Postgraduate School
Monterey, CA 93943

4. Dr. Nelson D. Ludlow, Code CS/Lw 2
Computer Science Department
Naval Postgraduate School
Monterey, CA 93943

5. CPT Kenneth H.Fritzsche...5
US Army PERSCOM
ATTN: PERSINSD
Alexandria, VA 22332

6. ECJ6-NP...1
HQ USEUCOM
Unit 30400 Box 1000
APO AE 09128